Regulatory Toxicology and Pharmacology 73 (2015) 770–779

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

A global initiative to refine acute inhalation studies through the use of 'evident toxicity' as an endpoint: Towards adoption of the fixed concentration procedure

Fiona Sewell [a, *], Ian Ragan [b], Tim Marczylo [c], Brian Anderson [d], Anne Braun [e], Warren Casey [f], Ngaire Dennison [g], David Griffiths [h], Robert Guest [h], Tom Holmes [i], Ton van Huygevoort [j], Ian Indans [k], Terry Kenny [l], Hajime Kojima [m], Kyuhong Lee [n], Pilar Prieto [o], Paul Smith [p], Jason Smedley [q], William S. Stokes [r], Gary Wnorowski [s], Graham Horgan [t]

[a] NC3Rs, UK
[b] Board Member, NC3Rs, UK
[c] Public Health England, UK
[d] Harlan Laboratories, Switzerland
[e] INERIS, France
[f] NICEATM, USA
[g] Home Office, UK
[h] Harlan Laboratories, UK
[i] Exponent International Limited, UK
[j] WIL Research, The Netherlands
[k] Health and Safety Executive, UK
[l] Huntingdon Life Sciences, UK
[m] JaCVAM (Japanese Center for the Validation of Alternative Methods), NIHS (National Institute of Health Sciences), Japan
[n] Korea Institute of Toxicology, South Korea
[o] EURL ECVAM, Systems Toxicology Unit, Institute for Health and Consumer Protection, European Commission, Joint Research Centre, Ispra, Italy
[p] Charles River Laboratories, Edinburgh, UK
[q] Charles River Laboratories, OH, USA
[r] U.S. Department of Agriculture, Animal and Plant Health Inspection Service, Animal Care, USA
[s] Product Safety Laboratories, USA
[t] BioSS, UK

ARTICLE INFO

Article history:
Received 16 June 2015
Received in revised form
19 October 2015
Accepted 20 October 2015
Available online 23 October 2015

Keywords:
Acute inhalation studies
3Rs
Evident toxicity
Fixed concentration procedure (FCP)
Refinement
Regulatory toxicology
TG4303
TG436
TG433

ABSTRACT

Acute inhalation studies are conducted in animals as part of chemical hazard identification and characterisation, including for classification and labelling purposes. Current accepted methods use death as an endpoint (OECD TG403 and TG436), whereas the fixed concentration procedure (FCP) (draft OECD TG433) uses fewer animals and replaces lethality as an endpoint with 'evident toxicity.' Evident toxicity is defined as clear signs of toxicity that predict exposure to the next highest concentration will cause severe toxicity or death in most animals. A global initiative including 20 organisations, led by the National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs) has shared data on the clinical signs recorded during acute inhalation studies for 172 substances (primarily dusts or mists) with the aim of making evident toxicity more objective and transferable between laboratories. Pairs of studies (5 male or 5 female rats) with at least a two-fold change in concentration were analysed to determine if there are any signs at the lower dose that could have predicted severe toxicity or death at the higher concentration. The results show that signs such as body weight loss (>10% pre-dosing weight), irregular respiration, tremors and hypoactivity, seen at least once in at least one animal after the day of dosing are highly predictive (positive predictive value > 90%) of severe toxicity or death at the next

Abbreviations: fixed concentration procedure, FCP.

* Corresponding author.
E-mail address: fiona.sewell@nc3rs.org.uk (F. Sewell).

http://dx.doi.org/10.1016/j.yrtph.2015.10.018

highest concentration. The working group has used these data to propose changes to TG433 that incorporate a clear indication of the clinical signs that define evident toxicity.

1. Introduction

1.1. Background

Acute inhalation studies are conducted in animals as part of chemical hazard identification and characterisation. Current accepted methods, LC50 (OECD TG403 (OECD, 2009a)) and the acute toxic class (ATC) (TG436 (OECD, 2009b)) use death as an endpoint. These are described in more detail below. In an effort to reduce animal numbers and to improve welfare, an alternative fixed concentration procedure (FCP) was proposed in 2004 (draft OECD TG433 (OECD, 2004) which replaced lethality as an endpoint with 'evident toxicity.' This was defined as those signs of toxicity that predict severe toxicity or death in most animals at the next highest concentration of the chemical. The FCP was dropped from the OECD work plan in 2007 because of a lack of evidence for comparable performance with TG403 and TG436, suspected sex differences in the level of toxic effects (since the FCP was originally proposed to use females as the default sex) and the ill-defined and subjective nature of evident toxicity. The first two issues have been resolved (Price et al., 2011; Stallard et al., 2011) through work supported by the UK National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs) thereby leaving only the definition of evident toxicity to be determined. To this end, the NC3Rs launched a global initiative involving 20 organisations with the aim of making evident toxicity more objective and transferable between laboratories. The group shared data on the clinical signs recorded during acute inhalation studies for 172 substances, the majority of which fell under the category of dusts and mists (from completed studies held in the archives of participating laboratories), and determined which signs have high positive predictive value (PPV) for severe toxicity or death at the next highest concentration (as described below). The draft OECD TG433 is now back on the OECD work plan, pending the outcome of this work.

1.2. Acute inhalation studies

The two existing guidelines (TG403 and TG436) are described here in some detail because the data used for the analysis in this paper originated from studies run according to these protocols. The FCP (draft TG433) is the preferable method for investigation of acute inhalation toxicity for classification and labelling purposes based on animal welfare grounds (preventing unnecessary suffering by eliminating the need to test at higher actual lethal doses). This method has been shown to be comparable with both existing methods in estimating the toxic class to which a substance belongs (Stallard et al., 2011).

Table 1
GHS classifications for LC_{50} by inhalation.

GHS category	Vapours (mg/L)	Dusts and mists (mg/L)	Gases (ppm)
1 (most toxic)	≤0.5	≤0.05	≤100
2	>0.5 and ≤2	>0.05 and ≤0.5	>100 and ≤500
3	>2 and ≤10	>0.5 and ≤1	>500 and ≤2,500
4	>10 and ≤20	>1 and ≤5	>2,500 and ≤20,000
5 (least toxic)	>20	>5	>20,000

GHS, Globally Harmonised System; LC_{50}, median concentration; ppm, parts per million.

1.2.1. LC_{50} method (TG403)

The LC_{50} of a substance is the concentration that can be expected to cause death in 50% of the animal population, where 'death' is defined as compound-related mortality within 14 days. The LC_{50} is used to classify substances (dust and mists, vapours, and gases) under the Globally Harmonised System of Classification and labelling of chemicals (GHS) (OECD, 2001). The test specifies that 10 animals (5 males and 5 females) should be exposed at each of three concentration levels. The concentration levels should be sufficiently spaced to enable construction of a mortality curve and an estimate of the LC_{50} to be obtained. The LC_{50} is then used to classify the toxicity of the chemical, according to Table 1 and as illustrated in Fig. 1.

1.2.2. Acute toxic class method (TG436)

The acute toxic class (ATC) method (TG436) has been accepted as an alternative method to the LC_{50} test (OECD TG403). Whilst the test uses fewer animals, death is still used as an endpoint. The test specifies that 6 animals (3 males and 3 females) are tested at fixed concentrations that form the upper limit of the GHS categories (e.g. 0.05, 0.5, 1 and 5 mg/L for dusts and mists) (Table 1). The starting concentration is either the highest concentration, or that which is expected to lead to mortality in some of the exposed animals, based on prior information. At each concentration decisions are based on the number of observed deaths from the combined group of animals. Either a classification is made or testing continues at the next higher or lower concentration, depending on the starting concentration, as shown in Fig. 2.

1.2.3. The fixed concentration procedure (FCP) (TG433)

The FCP test method is similar to the ATC method above but decisions and classifications are instead based on evident toxicity – clear signs of toxicity such that it can be predicted that exposure to the next highest concentration would cause death in most animals. The draft FCP protocol starts with a sighting study in which single female animals are exposed sequentially to one or more concentrations. Information from the sighting study can be used to classify the substance (if there is death at the lowest concentration the substance is classified into the most toxic class) or to guide decisions for an appropriate starting concentration of the main study. Comparison of the FCP test with the existing methods showed that, in the absence of sex differences, the results are similar (Price et al., 2011). Since the original FCP design proposed testing in female rats, the NC3Rs working group suggested the inclusion of a modified sighting study to take into account any sex differences in sensitivity. This involves the testing of one male and one female, to choose the most sensitive gender to take forward to main study testing. The main study then uses females (unless males are indicated as the more sensitive sex), where groups of five animals are exposed at each concentration until a decision on classification can be made. As for the ATC, substances are tested at fixed concentrations that form the upper limit of the GHS categories (e.g. 0.05, 0.5, 1 and 5 mg/l for dusts and mists) (Table 1). At each concentration decisions are based on the number of deaths and/or the number of animals experiencing evident toxicity, and either a classification is made or testing continues at the next higher or lower concentration, depending on the starting concentration (Fig. 3).

An issue that needed to be addressed by the group is the

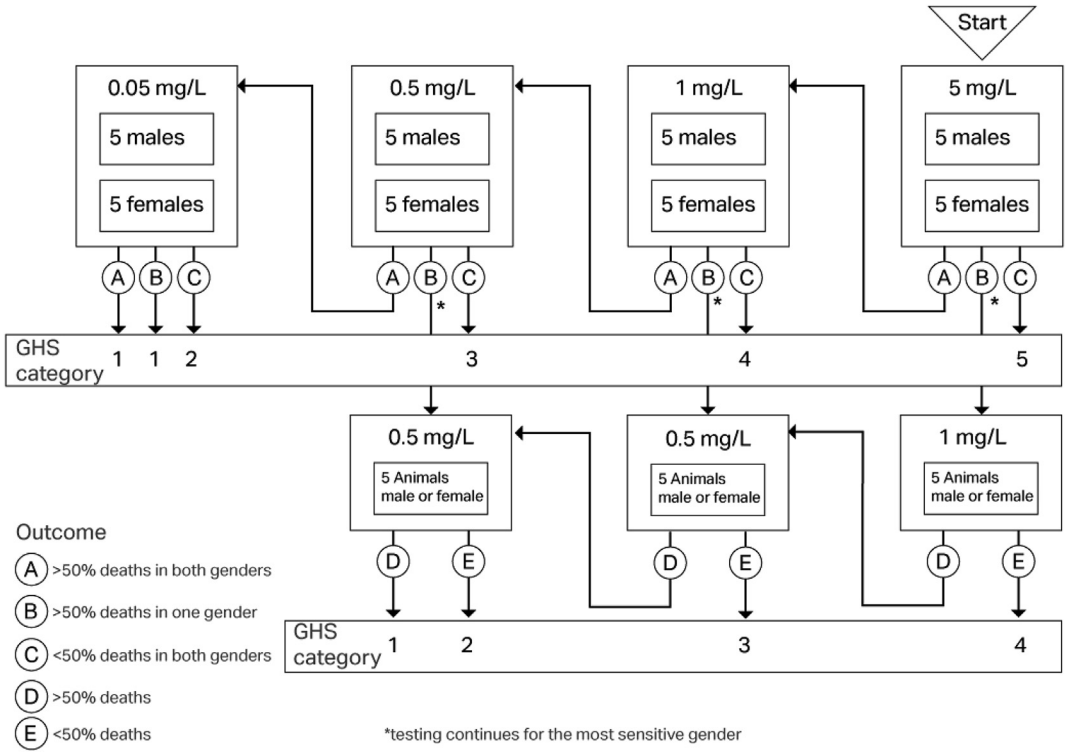

Fig. 1. LC_{50} test (OECD test guideline 403) for dusts and mists, using example concentrations, starting at 5 mg/L (Price et al., 2010). Please note the LC_{50} test method does not require fixed concentrations, but specifies that 10 animals (5 males and 5 females) should be exposed at three different concentration levels. The concentration levels should be sufficiently spaced to enable construction of a mortality curve so that an estimation of the LC_{50} can be obtained.

unequal interval between concentrations which varies from 2 to 10 fold within a class of substance (Table 1). The impact of this on the determination of evident toxicity is discussed below.

1.3. NC3Rs FCP working group and objectives

The NC3Rs assembled a working group of individuals from 19 other organisations listed as co-authors of this paper. The organisations include contract research organisations (CROs) and regulatory and standards bodies from Europe, US, Korea and Japan. The ultimate objective of the group is to encourage the adoption of the FCP as a preferred alternative to existing methods (TG403 and TG436) for acute inhalation studies. This is to be achieved through the generation of an evidence base of clinical signs to identify signs that predict exposure to a higher concentration would result in severe toxicity or death. The development of objective and validated assessment criteria for evident toxicity will provide guidance on the recognition of evident toxicity to inform decisions made during a study and for classification and labelling purposes. The raw data for the analysis were provided by six of the collaborating

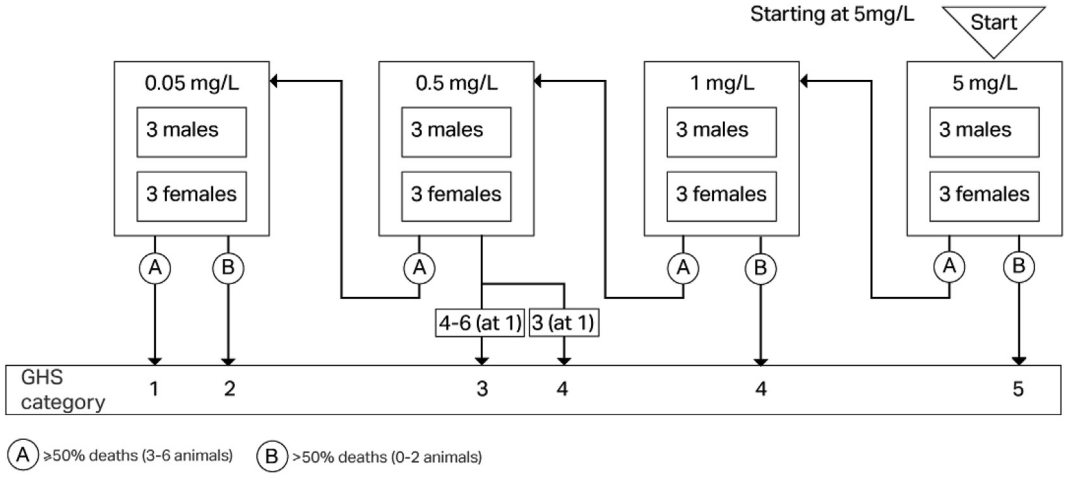

Fig. 2. Acute toxic class (ATC) method for dusts and mists for an example starting concentration of 5 mg/L (Price et al., 2010). Please note, the ATC method specifies that 6 animals (3 males and 3 females) are tested at fixed concentrations that form the upper limit of the GHS categories. The starting concentration is either the highest concentration, or that which is expected to lead to mortality in some of the exposed animals, based on prior information.

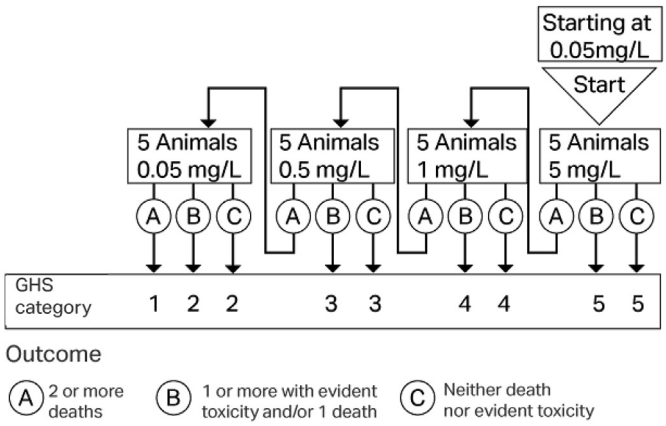

Fig. 3. Fixed concentration procedure (FCP) method for dusts and mists for an example starting concentration of 5 mg/L (Price et al., 2010). Please note, the draft test guideline specifies that substances are tested at fixed concentrations that form the upper limit of the GHS categories. The starting concentration is chosen to be the fixed concentration level that is most likely to lead to evident toxicity but not death.

organisations after initial agreement on the coding and recording of clinical signs. The data were received by the NC3Rs and anonymised to ensure that data could not be linked back to the originating organisation or to the chemical being tested.

2. Methods

2.1. Clinical signs recording system and data collection

A clinical signs recording system was developed with the expertise of the working group and piloted at four UK CROs to record the clinical signs observed in individual animals during acute inhalation studies. A code was assigned to each clinical sign observed. The trial was extended to collect data on the clinical signs observed in individual animals during acute inhalation studies conducted at six international laboratories for classification and labelling purposes. Table 2 lists the clinical signs and codes included in the clinical signs lexicon.

Data on the clinical signs observed in individual animals in acute inhalation studies for 172 substances, carried out at two or more concentrations, were provided by six international laboratories. These data were from either LC_{50} or ATC tests, but since the monitoring procedure is the same as for the FCP (i.e. daily monitoring of clinical signs for 14 days) they could be used to evaluate evident toxicity relevant for the FCP protocol. However, as detailed below, various exclusion criteria were applied to the data to make them more relevant to the FCP. Information was also collected on the type of substance (gas, dust, or vapour), as well as the test performed (LC_{50} or ATC), and the outcome of the study (e.g. further testing or GHS classification) (Table 3). Although listing the severity of each clinical sign was an option in the survey, this was not recorded systematically, and, with the exception of body weight loss, for the sake of consistency was not further considered in the analysis of evident toxicity.

2.2. Harmonisation of the dataset

Since archived data were collected from a number of laboratories there was some variation in the original terminology and signs recorded. Consequently a retrospective, harmonised list of clinical signs was agreed upon by all members of the working group for the purposes of analysis, and appropriate re-coding was made where needed. For example rales (dry) and rales (moist) were reassigned to code RN (noisy respiration). Varying descriptions of gait (shuffling, splayed, tiptoe) were grouped and assigned code AG (abnormal gait). Loose faeces and diarrhoea were deemed equivalent.

Table 2
Clinical sign lexicon.

General		***Movement***		***Respiration***	
BW	Body weight loss	**A**	Ataxia	**IR**	Irregular respiration
CTT	Cold to touch/hypothermia	**AG**	Abnormal gait	**RD**	Slow respiration
EM	Thin	**C**	Circling movement	**RG**	Gasping
DH	Dehydration	**ET**	Elevated tail	**RI**	Rapid/fast respiration
WTT	Warm to touch	**HT**	Tilted head	**RL**	Laboured respiration/dyspnea
Behaviour		**LL**	Limited use of limbs	**RN**	Noisy respiration
AGG	Aggressive	**RM**	RS SR	**RS**	Sneezing
AR	Appetite reduced	**RR**	Impaired righting reflex	**SR**	Shallow respiration
HS	Heightened sensitivity to sound	**WR**	Writhing	***Secretion/excretion***	
HST	Heightened sensitivity to touch	***Tremors/convulsions***		**DU**	Diuresis
I	Overactive/hyperactive	**CV**	Convulsions	**FAC**	Faeces abnormal colour
L	Underactive/hypoactive	**T**	Body tremors	**FR**	Faeces reduced
MU	Self-mutilation	**TW**	Twitching/fasciculation	**LF**	Loose faeces/diarrhoea
Q	Subdued	***Posture/muscle tone***		**ND**	Increased ocular–nasal lacrimation
V	Vocalising	**H**	Hunched posture	**OD**	Oral discharge
Appearance		**IBT**	Increased body tone	**UR**	Urine abnormal colour
AD	Abdominal discomfort	**PO**	Prone/flat posture	**S**	Increased salivation
CY	Cyanosis	**PR**	Prostrate	***Staining***	
DA	Distended abdomen	**RBT**	Reduced body tone	**FAS**	Facial staining
DF	Distended/swelling face	***Eyes***		**FS**	Generalised fur staining (caused by test item)
ES	Superficial eschar (scab)	**B**	Eye bulging	**SH**	Stained head
HL	Hair loss/alopecia	**CP**	Constricted pupil	**SF**	Generalised brown fur staining
P	Pilo-erection	**DE**	Eyes dull	**UGS**	Ano-genital staining
PA	Pallor	**DP**	Dilated pupil	***Consciousness***	
PP	Prolapsed penis	**EC**	Eye closed	**CO**	Unconscious
SKT	Skin tenting	**ED**	Eye damaged	**MO**	Moribund
TB	Teeth broken	**OE**	Opacity of eye	**O**	Active & healthy
UG	Rough/unkempt coat	**PtPC**	Eyes partially closed/ptosis	**XD**	Found dead
WF	Wet coat	**SW**	Eyelids swollen	**XE**	Animal euthanized after observations

Table 3
Substance information sheet.

Substance information sheet		Please provide additional information where possible
Inhalation type	[Select]	Please select
Method of inhalation	[Select]	Please select
Particle size (mmad)	[Enter]	
Concentration tested	[Enter]	Please specify units
Start time of exposure	[Enter]	
Duration of exposure	[Enter]	
Species tested	[Enter]	
Strain tested	[Enter]	
Chemical class	[Enter]	
Known toxic effects or mode(s) of toxicity	[Enter]	
Subsequent GHS classification *(please select)*	[Select]	Please select
Outcome of study	[Select]	Please select

2.3. Data analysis

The data were analysed to determine whether there were any clinical signs (or combination of signs) observed in animals at a lower concentration that could predict toxicity at the next or higher concentration. Although evident toxicity is defined in the guideline as that which 'predicts deaths in a majority of animals at the next higher concentration', thereby enabling classification according to Outcome B in Fig. 3, toxicity enabling classification according to Outcome A specifies only 2 or more deaths per group of 5, i.e. 40% or more. In view of this inconsistency, we agreed to define evident toxicity as that which predicts 'toxicity' at the higher concentration, itself defined as death, or severe toxicity requiring euthanasia, in two or more out of five animals.

Since the dataset included studies that were conducted according to both the LC_{50} and ATC protocols, there was wide variation in the concentration levels and intervals used. The fold changes in concentrations for the FCP protocol range from two to ten-fold, depending on the substance class, while the fold changes in concentrations in the data set varied from 1.02 to 20-fold. Therefore in this analysis tests were restricted to those with a two or greater fold difference in concentration between the lower and the higher exposure (Fig. S1).

The parameters used to analyse the dataset are the positive predictive values (PPV), the specificity, the sensitivity and the false positive or negative rates where:

- PPV is defined as the proportion of times that the presence of the clinical sign at the lower concentration is predictive of toxicity at the next highest concentration
- Sensitivity is defined as the proportion of 'toxicities' (death of two or more animals) at the higher concentration that are associated with the clinical sign at the lower concentration
- Specificity is defined as the proportion of non-toxicity that is associated with the absence of the clinical sign at the lower concentration
- The false positive rate is the proportion of times that the presence of the clinical sign at the lower concentration does not predict toxicity at the next highest concentration (therefore 1-PPV)
- The false negative rate is the proportion of the times that the absence of the clinical sign at the lower concentration is associated with toxicity at the next highest concentration

In the context of the FCP test, high values of PPV are required to avoid false positives that would lead to a more severe (but incorrect) GHS classification. Admittedly this would err on the side of caution for human safety and 3Rs considerations, but would not be attractive from the business perspective. Sensitivity is less important because there is no expectation that a single sign would suffice to predict all higher concentration toxicities. In a purely 3Rs context, any sign with very high PPV is attractive regardless of the sensitivity, but very low levels of sensitivity are less useful in practice because of their rare occurrence and small contribution to determining evident toxicity. Lower sensitivity is related to a higher proportion of false negatives i.e. toxicity at the higher concentration not predicted because the sign was absent at the lower concentration. Since this absence would naturally lead to further testing at the higher concentration, false negatives do not compromise the assessment of hazard, but are unsatisfactory from the 3Rs perspective.

In order to relate the analysis to the draft FCP protocol, various inclusion criteria and definitions were applied to the data. Comparisons of data from a lower and higher concentration were restricted to:

- a single sex, either male or female (FCP proposes testing of females only unless males are indicated as the most sensitive sex in the sighting study)
- a difference in concentration of at least two-fold (as described earlier to reflect the minimum concentration change in the protocol, Table 1)
- studies that had 5 animals per sex per group (the FCP tests 5 males or females)
- signs that were observed on days other than day 0 (the day of dosing) so that signs could be definitively related to the test substance rather than the inhalation or restraint procedure and so that predictions were based on signs that persisted/appeared 24 h after exposure (considered in more detail below)
- signs observed up to 14 days after dosing (the FCP requires daily monitoring for 14 days after exposure)

As in the draft FCP protocol, toxicity at the higher concentration is defined as severe toxicity requiring euthanasia or death in at least two out of the five animals. Therefore data for animals found dead (XD) or euthanised after observation (XE) were combined or used interchangeably.

The impact on the results of not including these restrictions was considered and is referred to in the results section that follows.

2.4. Statistical analysis

The ability of signs at the lower concentration to predict toxicity at the higher concentration was determined by examining 2×2 tables of whether one or more lower concentration animals showed a sign (or combination of signs) and whether 2 or more higher concentration animals died or were euthanized, and calculating the proportion of correct predictions. Confidence intervals were obtained from the mid-P confidence interval adaptation of the Clopper-Pearson interval (Agresti and Gottard, 2005). Differences between prevalence of signs in male and female animals was tested by fitting generalised linear models for the number of animals out of 5 showing a sign, with terms for test and sex.

3. Results

3.1. Properties of the original dataset

The original data set of tests on 188 substances included 511 pairs of studies on 4638 animals. These included information from a variety of procedures (LC_{50} and ATC) and concentrations tested,

depending on the type of substance tested. All studies were carried out in rat (typically 5 males and 5 females, although there were a small number of studies with fewer or greater numbers of animals per group) and animals were usually exposed to the inhalation substance for 4 h and then monitored daily for 14 days. In studies conducted for a longer period (21 or 28 days), only the data from the first 14 days was included in the analysis. Studies that used fewer than 5 males or 5 females were excluded from further analysis, as were data within studies in which the fold-change in exposure from one concentration to the next higher was less than two. Most of the remaining studies used both males and females although a number were conducted only in a single sex. All of these were considered eligible and in the rest of this paper, we use the term "study" to indicate a set of data from either 5 males or 5 females at two concentrations differing by at least a factor of 2. There were 427 pairs of these studies, from 172 substances, involving 3695 animals. The majority of these substances fall under the category of dusts and mists (165 substances), with a small number of gases (5 substances) and vapours (2 substances). However, since the purpose of the exercise is to look at the clinical signs observed and link these to the prediction of death or severe toxicity at a higher dose, the class of compound is largely irrelevant.

3.2. Death as a predictor of toxicity at the next highest concentration

A large proportion of animals were found dead during the studies and some required euthanasia. There were only 44 studies in which there were no deaths or euthanasia at either concentration, 1 in which one or more deaths were found only at the lower concentration, 224 in which death was found only at the higher concentration and 158 in which death was found at both concentrations.

'Toxicity' was defined as the death of two or more animals at the higher concentration. The presence of two deaths per five animals at the lower concentration was not surprisingly strongly associated with toxicity at the next higher exposure with a PPV of 98%. On rare occasions, death was more frequent at the lower concentration for unknown reasons.

One death only (1/5 animals) recorded at the lower exposure was also highly predictive of toxicity at the higher concentration (PPV 93%, 95% CI 84–98%). A subset of studies (28 with male rats, 29 with females) was used to investigate whether the presence of additional signs in a group of animals in which one death had been recorded at the lower concentration could increase the PPV. Table 4 shows that a small number of additional signs observed at least once in one animal increased the PPV to 100% with varying degrees of sensitivity. Most of these had high predictive value in the absence of death as described later.

Finally, even when no deaths were recorded at the lower exposure, toxicity at the higher concentration occurred in 77% (95% CI 72–82%) of studies.

3.3. Signs observed on day 0

Signs observed on day 0 may have resulted from the restraint and/or inhalation procedure rather than the chemical itself. In fact some signs were only (e.g. wet coat, writhing) or mainly (e.g. bulging eyes, eyes closed, overactive) observed on day 0 lending weight to this argument. However, some of the most common and severe signs were seen on day 0 as well as on other days, and an analysis was carried out to see the effect of inclusion and exclusion of these signs. On the whole, inclusion of day 0 observations decreased the predictivity of the sign. In other words, signs that persisted 24 h after exposure improved the predictivity. A good example is hypoactivity (L), which was much more frequently observed on day 0 than on other days. However, as a predictor of toxicity, inclusion of the day 0 incidence decreased predictivity from 100% to 92% and specificity from 100% to 88% (see Table 5). Other signs such as irregular respiration (IR) were similarly affected but to a lesser extent (Table 5). The decreases in predictivity and specificity were adopted as the justification for excluding day 0 observations from our analyses. However, it should be noted that if severe signs are seen on day 0, the usual procedures should be followed to stop a study or euthanize as appropriate.

3.4. Final dataset for analysis

The full dataset with the properties listed in 2.3 consisted of 427 pairs of studies. Of these, 268 pairs of studies had no deaths at the lower concentration, and it is this subset which is the subject of the following analysis.

Fig. S1 shows the concentration ratios for the 268 pairs of studies included in the final dataset for analysis. The majority of pairs had a concentration ratio in the range of >2 to ≤5 (80%), with a large proportion of studies with a concentration ratio of >2 to ≤3 (39%). A smaller proportion of studies had concentration ratios of 5 or more (20%). Concentration ratios in excess of 10 were seen less frequently (8%).

3.5. Most common signs excluding death at the lower concentration

There was a wide range of clinical signs recorded, relating to behaviour, posture, appearance, secretions/excretions or respiratory related phenomena. The 20 most common signs, observed at least once in at least one animal from day 1 onwards, are shown in Table 6. The most common signs have on the whole higher values of sensitivity, as expected, but can have very variable PPVs and specificities. The effect of changing the number of animals experiencing a sign is shown in Table 7 for irregular respiration, for tests

Table 4
A list of a small number of additional clinical signs, that if observed in the remaining animals in the lower dose group where one animal has died, can increase the PPV. The death of one animal at the lower dose has a PPV of 93%.

Clinical sign	Code	PPV (95% CI)		Sensitivity (95% CI)		Specificity (95% CI)		No. studies	No. animals
Irregular respiration	IR	100.0	(90.2–100.0)	54.7	(41.3–67.7)	100.0	(47.3–100.0)	29	132
Abnormal gait	AG	100.0	(76.2–100.0)	20.8	(11.5–33.2)	100.0	(47.3–100.0)	11	25
Slow respiration	RD	100.0	(71.7–100.0)	17.0	(8.7–28.9)	100.0	(47.3–100.0)	9	32
Tremors	T	100.0	(65.2–100.0)	13.2	(6.0–24.4)	100.0	(47.3–100.0)	7	20
Oral discharge	OD	100.0	(60.7–100.0)	11.3	(4.8–22.1)	100.0	(47.3–100.0)	6	20
Hypoactivity	L	94.7	(83.7–99.1)	67.9	(54.6–79.4)	50.0	(9.5–90.6)	38	121
Bodyweight loss	BW	93.3	(79.7–98.9)	52.8	(39.5–65.9)	50.0	(9.5–90.6)	30	93
Hunched posture	H	93.1	(79.1–98.8)	50.9	(37.7–64.2)	50.0	(9.5–90.6)	29	87

Table 5
Effect of inclusion or exclusion of day 0 signs on PPV and specificity for irregular respiration (IR) and hypoactivity (L).

Clinical sign	Code		PPV (95% CI)		Sensitivity (95% CI)		Specificity (95% CI)	
Irregular respiration	IR	Including day 0	86.3	(78.3–92.2)	39.6	(33.2–46.4)	78.7	(67.1–87.6)
		Excluding day 0	89.0	(80.9–94.5)	35.3	(29.0–42.0)	85.2	(74.7–92.5)
Hypoactivity	L	Including day 0	92.3	(85.4–96.6)	40.6	(34.1–47.4)	88.5	(78.7–94.8)
		Excluding day 0	100.0	(92.4–100.0)	18.4	(13.6–24.1)	100.0	(95.2–100.0)

Table 6
The 20 most common signs, observed at least once in at least one animal from day 1 onwards. The table also shows the number of animals showing the sign at least once.

	Clinical sign	Code	No. animals
1	Irregular respiration	IR	325
2	Noisy respiration	RN	278
3	Hunched posture	H	237
4	Respiration increased	RI	169
5	Bodyweight loss	BW	163
6	Laboured respiration	RL	160
7	Piloerection	P	116
8	Faeces reduced	FR	107
9	Body staining	ST	99
10	Naso-ocular discharge	ND	91
11	Hypoactivity	L	87
12	Congested respiration	RC	87
13	Facial staining	FAS	56
14	Ano-genital staining	UGS	51
15	Appetite reduced	AR	33
16	Gasping	RG	30
17	Unkempt	UG	25
18	Tremors	T	15
19	Repetitive movements	RM	10
20	Distended abdomen	DA	8

in which both 5 males and 5 females were included. PPV and specificity are largely unchanged, but sensitivity decreases as the requirement to see the sign in a higher and higher proportion of the animals becomes increasingly unlikely. Similar conclusions were reached for other common signs examined, i.e. small changes to PPV and specificity and a loss of sensitivity. Therefore subsequent analyses and recommendations are based on observation of the sign at least once in at least one animal.

3.6. Clinical signs as predictors of toxicity at next highest concentration

Since toxicity occurred at the higher concentration in 77% of the studies where there were no deaths at the lower concentration, the determination of evident toxicity (as opposed to death as an endpoint) is only potentially useful if its predictive power (PPV) is greater than this value (77%) and comparable to that of the observation of 1 death in a group of 5 animals (93%). There were a number of individual signs whose PPVs were greater than 77% and a smaller number whose PPVs were comparable to that of 1 death and for which the lower 95% confidence limits also exceeded 77% (Table 8). These were hypoactivity (L), tremors (T), body weight loss (BW) and irregular respiration (IR). Tremors were not commonly observed (low sensitivity) and therefore the contribution of this sign is small.

3.7. Body weight loss

In view of the possible value of body weight loss as an objective marker of evident toxicity, we examined whether further refinement could be achieved by consideration of the extent of loss. In the lexicon of signs, there were a number of sub-categories whose numbers were pooled in the analysis of Table 8. These were: EM (thin appearance, 252 observations), BW (extent not specified, 1201 observations), BW (mild) (reduced weight gain, 182 observations), BW (moderate) (loss 10–20%, 941 observations), BW (substantial) (loss >20%, 144 observations). Where losses are quantified, these are percentage lost compared to the pre-dosing weight on day 0. As shown in Table 9, there is little to be gained from use of these sub-categories as any modest improvement in PPV is at the expense of wider confidence limits and much reduced sensitivity.

3.8. Combinations and co-occurrence of signs

Test data were examined for the presence of combinations of signs i.e. one animal or more with either of sign A or sign B at the lower exposure. All possible combinations of 2 signs were considered but the gains in sensitivity were small because of the strong co-occurrence of signs i.e. signs are not truly independent of each other. Similarly, inclusion of a third or fourth sign had progressively lesser impact on any parameter.

The presence of any one of the three most highly predictive signs (hypoactivity, body weight loss and irregular respiration) in at least one animal had a PPV of nearly 91%, predicting 52% (sensitivity) of the toxicity that occurred at the higher concentration with only a 9% false positive rate (Table 10), which compares favourably with the false positive rate of a single death at the lower concentration. Addition of further signs in the same manner predicted a larger proportion of the toxicity at the higher concentration, but inevitably at the expense of a greater false positive rate.

3.9. The effect of varying concentration ratios

The interval between the high and the low concentration in the draft FCP guideline varies between two- and ten-fold. It was to be expected that the larger the concentration ratio the more likely it would be that toxicity would be found at the higher concentration and conversely that a small concentration ratio would be associated with greater risk of false positives. This was examined for the common signs by comparing, for each sign in Table 11, the average concentration ratio where false positive predictions were made with the average concentration ratio for true positive predictions.

Table 7
The effect of changing the number of animals experiencing clinical sign irregular respiration (IR) on predictivity (PPV).

No. animals with sign IR	PPV (95% CI)		Sensitivity (95% CI)		Specificity (95% CI)		No. studies
1	89.0	(80.9–94.5)	35.3	(29.0–42.0)	85.2	(74.7–92.5)	82
2	87.5	(78.3–93.7)	30.4	(24.5–36.9)	85.2	(74.7–92.5)	72
3	86.4	(76.5–93.1)	27.5	(21.8–33.9)	85.2	(74.7–92.5)	66
4	84.5	(73.5–92.1)	23.7	(18.3–29.8)	85.2	(74.7–92.5)	58
5	85.1	(72.8–93.2)	19.3	14.4–25.1)	88.5	(78.7–94.8)	47

Table 8
The highly predictive signs in the absence of death (PPV above 73%). Signs observed at least once in at least one animal, excluding day 0.

Clinical sign	Code	PPV (95% CI)		Sensitivity (95% CI)		Specificity (95% CI)		No. studies
Hypoactivity	L	100.0	(92.4–100.0)	18.4	(13.6–24.1)	100.0	(95.2–100.0)	38
Tremors	T	100.0	(68.8–100.0)	3.9	(1.9–7.2)	100.0	(95.2–100.0)	8
Bodyweight loss	BW	94.0	(84.6–98.4)	22.7	(17.4–28.8)	95.1	(87.2–98.7)	50
Irregular respiration	IR	89.0	(80.9–94.5)	35.3	(29.0–42.0)	85.2	(74.7–92.5)	82
Body staining	ST	88.5	(71.8–97.0)	11.1	(7.4–15.9)	95.1	(87.2–98.7)	26
Ano-genital staining	UGS	86.4	(67.3–96.4)	9.2	(5.8–13.7)	95.1	(87.2–98.7)	22
Faeces reduced	FR	85.3	(70.4–94.4)	14.0	(9.8–19.2)	91.8	(82.8–96.9)	34
Naso-ocular discharge	ND	85.0	(71.4–93.7)	16.4	(11.9–21.9)	90.2	(80.7–95.9)	40
Noisy respiration	RN	81.2	(71.9–88.4)	33.3	(27.2–40.0)	73.8	(61.7–83.6)	85
Hunched posture	H	78.8	(66.3–88.3)	19.8	(14.8–25.6)	82.0	(70.9–90.1)	52

Table 9
Effect of sub-categories of bodyweight loss on PPV. Sign observed at least once in at least one animal.

Clinical sign	Code	PPV (95% CI)		Sensitivity (95% CI)		Specificity (95% CI)		No. studies
Bodyweight loss (unspecified)	BW (unspecified)	100.0	(77.9–100.0)	5.8	(3.2–9.6)	100.0	(95.2–100.0)	12
Bodyweight loss (mild)[a]	BW (mild)	100.0	(36.8–100.0)	1.4	(0.4–3.9)	100.0	(3.9–95.2)	3
Bodyweight loss (moderate)[b]	BW (moderate)	94.4	(82.9–99.0)	16.4	(11.9–21.9)	96.7	(21.9–89.6)	36
Bodyweight loss (substantial)[c]	BW (substantial)	100.0	(22.4–100.0)	1.0	(0.2–3.1)	100.0	(3.1–95.2)	2
Thin	Thin	50.0	(2.5–97.5)	0.5	(0.1–2.3)	98.4	(2.3–92.2)	2

[a] Reduced weight gain.
[b] Weight loss 10–20% compared to pre-dosing weight on day 0.
[c] Weight loss >20% compared to pre-dosing weight on day 0.

Table 10
Combinations of highly predictive signs. PPV, sensitivity and specificity of studies where any one of the following signs are seen in any animal at least once, excluding day 0: bodyweight loss (BW), hypoactivity (L) and/or irregular respiration (IR).

Combination of clinical signs	Codes	PPV (95% CI)		Sensitivity (95% CI)		Specificity (95% CI)		No. studies
Hypoactivity, bodyweight loss and/or irregular respiration	L, BW and/or IR	90.7	(84.4–95.0)	51.7	(44.9–58.4)	82.0	(70.9–90.1)	118

As shown in the Table, false positive prediction was more frequently associated with smaller concentration ratios, an argument for a more rational experimental design. However, for the four signs of highest PPV (hypoactivity, tremors, bodyweight loss and irregular respiration), the first two (hypoactivity and tremors) are never associated with false positives, while for the second two (bodyweight loss and irregular respiration), there was no significant effect of concentration ratio.

3.10. Sex differences

Some signs were more prevalent in one sex than the other. Among the more common signs, ano-genital staining (UGS) was more prevalent in females ($p = 0.0002$) while facial staining (FAS) and gasping (RG) were marginally more common in males ($p = 0.028$ and 0.044, respectively). However, the predictivity of these signs did not differ between males and females even for UGS because of the wide confidence limits particularly for the sex with fewer observations where sensitivity was reduced.

4. Discussion of results

There are advantages and disadvantages of using archived data for analyses of this type. The advantage is the gathering of a large data set from which robust conclusions can be drawn. The disadvantage is that there is no way to control for laboratory-specific differences in the nomenclature and recording of clinical signs. We have tried to correct for this by creating a list of clinical signs

Table 11
Concentration ratios (the ratio between the lower and higher dose) and PPVs for commonly observed clinical signs. For each of the signs shown, average concentration ratios in those studies giving rise to false positive prediction of toxicity were compared with the average concentration ratios in those studies giving rise to true prediction of toxicity. The p-value is calculated from a t-test of whether the mean concentration ratio differs between false and true positives.

Clinical sign	Code	PPV (95% CI)		Concentration ratios		p-value
				False positive (sign present but toxicity does not occur)	True positive (sign present & toxicity occurs)	
Hypoactivity	L	100.0	(92.4–100.0)	–	4.09	–
Tremors	T	100.0	(68.8–100.0)	–	4.56	–
Bodyweight loss	BW	94.0	(84.6–98.4)	4.97	3.57	0.610
Irregular respiration	IR	89.0	(80.9–94.5)	4.30	5.14	0.296
Body staining	ST	88.5	(71.8–97.0)	2.06	3.61	0.001
Ano-genital staining	UGS	86.4	(67.3–96.4)	4.71	3.50	0.671
Faeces reduced	FR	85.3	(70.4–94.4)	2.07	3.89	<0.001
Naso-ocular discharge	ND	85.0	(71.4–93.7)	3.40	4.27	0.184
Noisy respiration	RN	81.2	(71.9–88.4)	3.62	3.90	0.597
Hunched posture	H	78.8	(66.3–88.3)	2.55	3.57	0.011

that was then piloted by four different CROs before applying it to the whole data set. Even so, there was some ambiguity and variation in detail in the responses that were provided (such as indications of severity) and some judgement had to be applied in the interests of harmonisation. Where any reclassification was proposed, this was always agreed with the originating organisation.

In the dataset that had no deaths at the lower concentration, and regardless of what clinical signs were noted, toxicity occurred in 77% of the studies at the next higher concentration. Therefore, the use of clinical signs other than death to define evident toxicity clearly has to have a PPV greater than this to have any potential value. What lower limit of the PPV above 77% is acceptable is determined by attitudes to false positive rates, and in this there is likely to be some diversity of opinion between safety, commercial and 3Rs perspectives. The other benchmark of utility is comparability with the predictive value of a single death at the lower concentration (PPV 93%), since this had already been proposed as a criterion for GHS classification in the draft OECD guideline. We consider therefore several levels of the utility of clinical signs for determining evident toxicity. Firstly there are signs whose PPVs are comparable to that of a single death, whose lower 95% confidence limits exceed 77%, and for which specificity is high and sensitivity is appreciable. These are L (hypoactivity), BW (body weight loss) and IR (irregular respiration). These three signs satisfy the need to avoid false positives, and are quite commonly observed resulting in appreciable sensitivity. Body weight loss has previously been shown to be a reliable objective marker for the determination of the maximum tolerated dose (MTD) in short term toxicity tests in animals to inform pharmaceutical development of potential candidate drugs (Chapman et al., 2013). Secondly, the presence of T (tremors) also satisfies the need to avoid false positives but it is less commonly observed with a sensitivity of only 3.8%. We propose that the observation of any of these four signs indicates evident toxicity. Thirdly, there is a further group of signs with PPVs in excess of 77%, but with lower values of the lower 95% confidence limit, whose utility is further compromised in some instances by lower specificity and limited sensitivity. It is therefore a matter of debate whether the presence of these signs would be accepted as evident toxicity.

In view of the complications inherent in using historical data, it would be ideal to collect a prospective dataset in which the clinical signs to be monitored had been agreed in advance, and include additional information on severity in order to strengthen the case for including the third category of signs as indicators of evident toxicity.

A few signs showed differences in prevalence between males and females. Earlier work had shown that in the case that males are more sensitive than females, the FCP performs less well, and for that reason a sighting study in both sexes was proposed (Stallard et al., 2011). The present data do not alter this conclusion, since the choice of sex only determines the frequency of certain signs not their predictive value.

5. Recommendations and conclusions

5.1. Revised FCP protocol

Since we aim to encourage the adoption and acceptance of the FCP by making the decision of evident toxicity more objective and transparent we want to ensure our recommendations are as simple and effective as possible. Concentration fold change, severity, timing, onset and duration of signs, the number of animals displaying a sign were all considered and analysed for their ability to predict toxicity at the next highest dose. However, it was found that these analyses could be simplified to give a few highly predictive signs that could indicate evident toxicity if observed at least once, in one animal after the day of exposure.

If one or more of the signs shown in Table 12 (T, L, BW and IR) are observed in one or more animals during an acute inhalation study there is no need to conduct a further study at a higher concentration, since this demonstrates that evident toxicity has been reached and testing at a higher concentration will likely result in death or severe suffering. Please note, we recommended that a limit of 10% body weight is used, since this was the lowest quantifiable limit in the dataset, and body weight loss in excess of 10% (from pre-dosing weight on day 0) was shown to be highly predictive of toxicity at the next highest concentration.

Therefore the recommendations for the FCP protocol are as follows:

- A sighting study using one male and one female rat at each concentration should be undertaken and subsequent testing conducted in the more sensitive sex, or in females if no difference in sensitivity between sexes is observed (Stallard et al., 2011).
- If 2 or more deaths occur in the main study, substance classification can be made or further testing should be carried out at the concentration level below that tested, depending on the starting concentration or previous concentrations tested (Fig. 3, Outcome A).
- If there is 1 death or there is evident toxicity, classification can be made according to Table 1, irrespective of the starting concentration or previous concentrations tested (Fig. 3, Outcome B). Evident toxicity has been reached if sign T, L, BW or IR is seen at least once in at least one animal from day 1. However, these recommendations are not intended to overrule study director experience and judgement. If severe toxicity is observed on the day of exposure or through the presentation of other signs not listed here, the study should be stopped and animals euthanized as appropriate.
- If there are no deaths and evident toxicity is not observed, non-classification can be concluded or further testing should be carried out at the concentration level above that tested, depending on the starting concentration or previous concentrations tested (Fig. 3, Outcome C).

5.2. Next steps and conclusions

Adoption of the FCP in preference to the currently accepted methods will refine acute inhalation studies through the use of evident toxicity rather than death as an endpoint. In addition, since the FCP uses fewer animals than currently accepted methods, it also has the potential to reduce the number of rats used in acute inhalation toxicology studies worldwide.

With all three areas of concern now addressed: (i) comparability; (ii) potential for sex differences; and (iii) the recognition of evident toxicity, the working group will now seek international

Table 12
Guidance on the recognition of evident toxicity.

Guidance on the recognition of evident toxicity	
Evident toxicity has been reached if one or more animals display any one of the listed signs (from day one onwards):	
Clinical sign	Code
Hypoactivity	L
Tremors	T
Bodyweight loss (>10%)	BW
Irregular respiration	IR

acceptance of the FCP through adoption of OECD TG433 and encourage its worldwide use in preference to other methods.

Disclaimer

This presentation does not necessarily reflect the official position of the USDA or any U.S. Agency.

Acknowledgements

There was no specific financial support for this paper other than the invested time of the authors. We would like to thank Dr Katie Lidster and Dr Joanna Edwards for their time spent coding the data.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.yrtph.2015.10.018.

Transparency document

Transparency document related to this article can be found online at http://dx.doi.org/10.1016/j.yrtph.2015.10.018.

References

Agresti, A., Gottard, A., 2005. Comment: randonmized confidence intervals and the mid-*P* approach. Stat. Sci. 20, 367–371.

Chapman, K., Sewell, F., Allais, L., Delongeas, J.L., Donald, E., Festag, M., Kervyn, S., Ockert, D., Nogues, V., Palmer, H., Popovic, M., Roosen, W., Schoenmakers, A., Somers, K., Stark, C., Stei, P., Robinson, S., 2013. A global pharmaceutical company initiative: an evidence-based approach to define the upper limit of body weight loss in short term toxicity studies. Regul. Toxicol. Pharmacol. – RTP 67, 27–38.

OECD, 2001. Harmonized Integrated Hazard Classification System for Human Health and Environmental Effects of Chemical Substances. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono%282001%296.

OECD, 2004. Draft Proposal for a New Guideline:433. http://www.oecd.org/chemicalsafety/testing/32035886.pdf.

OECD, 2009a. OECD Test Guideline 403: Acute Inhalation Toxicity. http://www.oecd-ilibrary.org/docserver/download/9740301e.pdf?expires=1433247636&id=id&accname=guest&checksum=7012504CE687B2E5614DB637989CE606.

OECD, 2009b. Test Guideline 436: Acute Inhalation Toxicity – Acute Toxic Class Method. http://www.oecd-ilibrary.org/docserver/download/9743601e.pdf?expires=1433247696&id=id&accname=guest&checksum=DF8991A8B9D88D13E75E914D1A2D5022.

Price, C., Stallard, N., Creton, S., Indans, I., Guest, R., Griffiths, D., Edwards, P., 2011. A statistical evaluation of the effects of gender differences in assessment of acute inhalation toxicity. Hum Exp Toxicol. 30, 217–238.

Stallard, N., Price, C., Creton, S., Indans, I., Guest, R., Griffiths, D., Edwards, P., 2011. A new sighting study for the fixed concentration procedure to allow for gender differences. Hum. Exp. Toxicol. 30, 239–249.